MR Images Classification Using Hybrid KNNSVM Algorithm

A. Kannan[1], V. Mohan[2], N. Anbazhagan[3]

[1]Department of MCA, K.L.N. College of Engineering, Sivagangai District, Tamilnadu, India

[2]Department of Mathematics, Thiagarajar College of Engineering, Madurai, Tamilnadu, India

[3]Department of Maths, Alagappa University, Karaikudi, Tamilnadu, India

[1]kannamca@yahoo.com; [2]vmohan@tce.edu; [3]anbazhagan_n@yahoo.co.in

Abstract

In the field of medical, image data play a vital role to assist the physicians in all kinds. Especially, Magnetic Resonance Image data will be very useful to diagnose brain tumors in human brain. Unfortunately, there are certain difficulties to classify those images to take sudden appropriate decisions to recover the identified disease. Hence, the concept of image mining is used to extract potential hidden information from the image data and those can be classified to take right decision for the early recover of the patient. Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) are very powerful and popular classification algorithms to classify image data. However, these two have their own drawbacks in certain situations. In this paper, both SVM and KNN have been merged to derive a hybrid KNNSVM algorithm to diagnose the MR Images in an effective manner with high accuracy rate and low error rate.

Key Words

GLCM; SVM; KNN; Texture

Introduction

Medical field produces huge number of image data to diagnose major diseases of human beings at regular intervals. At most, all data are in the form of digital image formats. Physicians are in a position to consult experts in order to take necessary remedial actions for the identified medical problems. To overcome this, there is a need to develop an image classification system which automatically examines the problem without assistance of any expertise. Though there are huge numbers of classification algorithms, there is a need to hybrid few of them to achieve good result with high reliability. Normally, the Bidirectional Associative Memory (BAM), Portable Neural Network (PNN), Support Vector Machine (SVM), Learning Vector Quantization (LVQ) and K-Nearest Neighbor (KNN) are popular image classification algorithms in the field of image mining techniques. BAM maps the patterns from an input layer X to patterns in an output layer Y, where each layer consists of a set of artificial neurons capable of representing the input or output patterns. BAM has the distinction that it is bidirectional, so it can also map the patterns from layer Y back to layer X. Though BAM has high accuracy rate, it consumes huge time to classify the data even for a lower number of images and also it consumes large memory spaces. The major task of the LVQ algorithm is to find the suitable optimal placement of feature vectors in input space in order to approximate the different class domains in the vector space where the training samples reside in. The input space is two dimensional. Typically, multiple feature vectors will be assigned to each class when attempting to model the class domains. The Probabilistic Neural Network (PNN), which has its origin in Specht's earlier work, has been recognized as a useful technique that can be used in classification problems. The PNN is designed by combining the techniques of kernel based estimator for the estimation of probability densities with the Bayes rule for the classification decision. The considerable advantages of the PNN classifier are its execution speed and the simplicity of the training process. In training process, a kernel function and its smoothing parameter will be selected. There is no need for weight adaptation as in other neural networks. The PNN and LVQ are the fastest execution algorithms, but produce low accuracy rate. SVM and KNN increase the accuracy rate and reduce error rate. However, both SVM and KNN have their own drawbacks. The efficiency and accuracy of the KNN algorithm will be low, when the numbers of samples are increased. i.e.) KNN-distance functions are anticipated to have poor classification performance as the dimensionality of the noisy data increases. In SVM, there is lower classification accuracy, if the sample data of the two classes in a

binary classification are all close to the separating hyper plane.

In this paper, a hybrid algorithm is designed by merging the concepts of the SVM and KNN classification algorithms to classify MR Images to conclude the results whether the stage of the tumor condition in human brain is 'Normal, 'Benign' or 'Malignant'. Based on the results, the physicians can take necessary remedial actions even without the consultant opinion of any expert. In this research, the conventional Sensitivity and Specificity measurements have been examined for the justification of the results.

Proposed Solution

In this work, supervised classification technique is applied to analyze MR Images. Supervised classification is a kind of process where known identified samples are classified to reach the targeted result. By this, the data set can be controlled by the analyst. In this process, it is very important to have desired classes and from there appropriate signatures can be formed effectively. The errors of the test image can easily be identified by examining the training set seriously. There are two major activities in this research work such as training process and testing process. In the training process, the MR images have been collected in the form of gray scale format. For this training set, MR Images have been taken from DICOM. The DICOM has already preprocessed and concluded certain categories of the images in a right way. However, the images are again to be preprocessed in order to improve the quality of the result. The Grey Level Co-occurrence Matrix (GLCM) will be calculated from those images in order to identify texture contents later. The GLCM will contain the information about the positions of pixels which have similar grey level values. A co-occurrence matrix will be a two-dimensional array 'P' in which the possible image values will be defined as rows and columns. Then, the texture features will be extracted from the collected stored supervised MR images (training) and those features will be kept in a database for future processes. The Table 2.1 shows the categories of training set for this paper. The Table 2.2 shows the 12 texture features that are extracted from the training set MR Images. However, the dimensionality of the features can considerably be reduced further to increase the speed of the processes.

The generated features of the training images will be optimized by using Sequence Forward Selection (SFS) method. The SFS is a common method used for feature selection. The features 'Cluster Prominence' and 'Entropy' have been received as the outcome optimized features after implementing SFS. Finally, the feature space will be formed based on these optimized features. In the testing process, the test image i.e., the patient's MR image will be received from the user and the GLCM features of that image will be computed and the texture features will be extracted in terms of feature vectors. Again, the 12 texture features will be extracted from the given test image. Of these, only the optimized features of the query image will be used for the test such as Entropy, Cluster Prominence etc. The purpose of choosing these 12 features is their high support to the core objective of this research work and sufficient elements to process MR Images. Based on these features, the KNN feature space can be formed. Finally, the hybrid algorithm will be applied to classify the given query (test) image. The subsequent steps and other relevant processes are illustrated in the Figure 2.1.

As mentioned earlier, initially, the test query MR Image is to be received from the user and its GLCM features have to be extracted. Then, the 12 features have to be optimized using SFS method to reduce features' dimensionality. Finally, the proposed hybridized KNNSVM algorithm is applied on the given query image. Initially, the KNN will be employed to identify whether the given query image falls in the category of 'Benign', 'Normal' or 'Malignant'. If the result is not concluded, then SVM1 is to be employed to identify the image either as 'Normal' or 'Abnormal'. If 'Normal', the result is concluded, otherwise, SVM2 will be employed to identify whether the stage of tumor is 'Benign' or 'Malignant'. The query image is to be compared with the existing results of training set.

The K-Nearest-Neighbour (KNN) algorithm measures the distance between a test sample and a set of training samples in the features space. Here, the training MR Images are supervised classification images and those images have already been labelled. The nearest neighbours for this test sample will be determined using distance measurement functions. Almost, every classification and clustering method needs a distance measure dist (q_i, tj) between the query

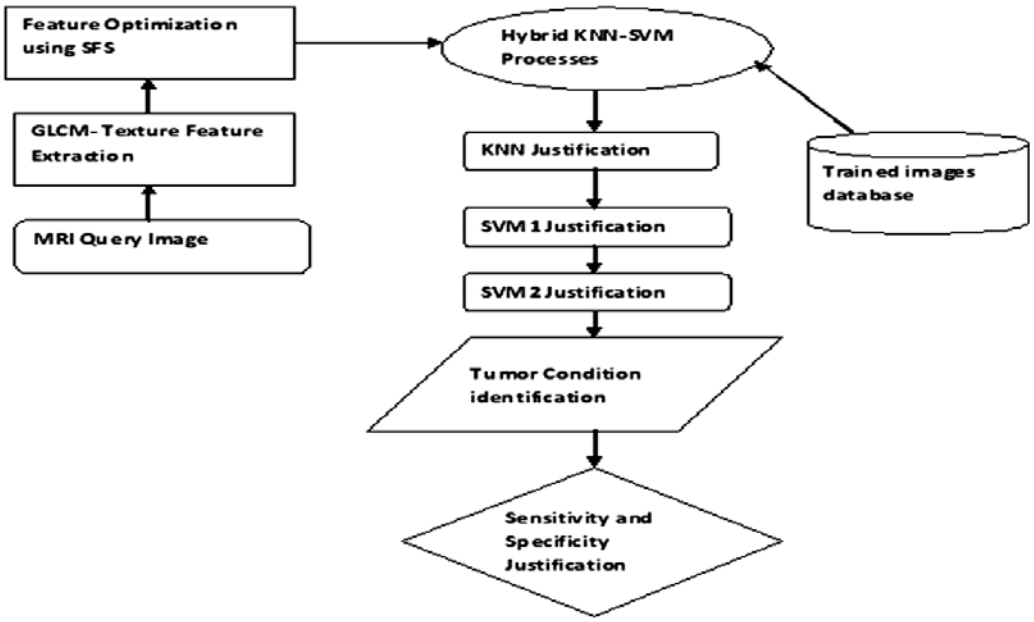

FIGURE 2.1 BLOCK DIAGRAM OF HYBRID KNN-SVM PROCESSES

sample and other samples. Common distance measurement methods are Euclidean or L2 distance, Mahalanobis distance, Manhattan or L1 distance or City Block distance methods. Here, the distances of the neighbours are calculated using Euclidean distance method, since the Euclidean distance method is very popular and simple method to calculate distance values between any pairs. Then, the neighbours distance values will be sorted out and the 'k' nearest neighbours will be picked out from that index.

The aim of the KNN classification is to obtain the nearest-neighbour list. Once the list is received, the query sample is classified based on the majority class of its nearest neighbours. The following Equation 2.1 is used to get majority vote for the given query image.

$$QC' = \underset{c}{\operatorname{argmax}} \sum_{(t_i, c_i) \in K_d} S(c = ci) \qquad (2.1)$$

Where 'c' is a class label, 't_i' is the training sample in i^{th} position, ''c_i' is the class label for the 'i^{th}' nearest neighbours and 'K_d' is the nearest neighbour list. The S(.) settling class function that returns the value 1 if it argument is true and 0 otherwise.

Hence, if the testing sample is same as the labels of the majority of its K-nearest neighbours, the test sample will be grouped to the category concerned of the classifications. Else, the current process will be switched over to SVM1. If still not concluded the result, the process of classification will be moved to SVM2.

TABLE 2.1 SUPERVISED MR IMAGES WITH LABEL

	Range of MR Image categories in training set		
	1-23	24-48	49-150
MRI Group	Normal	Benign	Malignant

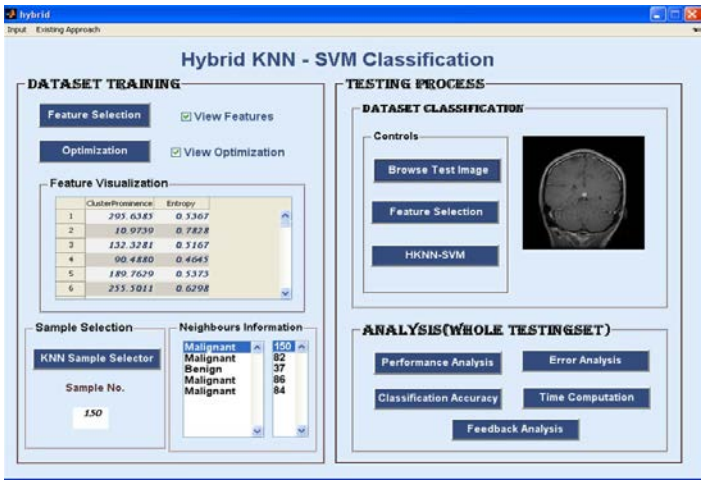

FIGURE 2.2 LOADING THE QUERY MRI IMAGE

TABLE 2.2 TEXTURE FEATURES OF TRAINING SAMPLES

S.No.	Contrast	Correlation	Clus.Prominence	Clus.Shade	Dissimilarity	Energy	Entropy	Homogenity	Homop	Max.Prob	Sosvh	Auto.Corre
1.	0.076978	0.955866	295.6385	30.12207	0.061946	0.834309	0.536679	0.971382	0.970524	0.913014	2.318267	2.317603
2.	0.166851	0.531982	10.97391	1.853608	0.15269	0.656248	0.7828	0.925562	0.925021	0.802967	1.453509	1.405617
3.	0.119304	0.856217	132.3281	13.22594	0.080142	0.826696	0.516733	0.965198	0.963728	0.908386	1.657289	1.632714
4.	0.087816	0.89289	90.48805	11.14623	0.060364	0.853755	0.464491	0.973639	0.972528	0.923576	1.659678	1.650949
5.	0.135047	0.898226	189.7629	20.72418	0.076582	0.828986	0.537348	0.968734	0.967142	0.909929	2.025196	1.994383
6.	0.166377	0.898959	255.5011	26.4797	0.091139	0.803471	0.629785	0.962908	0.96114	0.895728	2.277108	2.231843
7.	0.143117	0.906357	183.9123	21.68506	0.081487	0.802866	0.622724	0.96682	0.965064	0.895411	2.227421	2.193908
8.	0.152848	0.919698	294.6303	30.70965	0.093196	0.787922	0.672409	0.960849	0.959027	0.886946	2.48507	2.447587
9.	0.126741	0.918725	180.2494	21.49874	0.085601	0.78399	0.68227	0.962829	0.9612	0.884731	2.284015	2.259217
10.	0.098734	0.914628	78.51145	12.31779	0.067326	0.816186	0.56975	0.970778	0.969362	0.902888	2.004524	1.992722
11.	0.182278	0.906026	175.5874	23.39528	0.113449	0.73595	0.813125	0.952453	0.949936	0.856764	2.661172	2.610918
12.	0.115111	0.913163	68.13628	11.99739	0.078797	0.760548	0.703485	0.965531	0.96408	0.870807	2.238183	2.220095
13.	0.113291	0.931624	101.8365	16.34958	0.079589	0.747661	0.745594	0.965069	0.963542	0.863212	2.514127	2.498299
14.	0.16962	0.962881	634.0849	68.79708	0.11693	0.680062	0.933774	0.948824	0.946696	0.822429	4.693464	4.657397
15.	0.159652	0.96161	616.2992	65.40447	0.114399	0.705172	0.878222	0.949332	0.947265	0.838133	4.299983	4.26697
16.	0.16962	0.962881	634.0849	68.79708	0.11693	0.680062	0.933774	0.948824	0.946696	0.822429	4.693464	4.657397
17.	0.159652	0.96161	616.2992	65.40447	0.114399	0.705172	0.878222	0.949332	0.947265	0.838133	4.299983	4.26697
18.	0.13663	0.959732	498.4454	53.62866	0.1	0.734265	0.797713	0.955268	0.953618	0.855617	3.674399	3.650277
19.	0.112816	0.957732	427.2246	44.46549	0.087975	0.776689	0.695554	0.959712	0.95848	0.880538	3.057198	3.042049
20.	0.076978	0.955866	295.6385	30.12207	0.061946	0.834309	0.536679	0.971382	0.970524	0.913014	2.318267	2.317603
21.	0.159652	0.96161	616.2992	65.40447	0.114399	0.705172	0.878222	0.949332	0.947265	0.838133	4.299983	4.26697
22.	0.13663	0.959732	498.4454	53.62866	0.1	0.734265	0.797713	0.955268	0.953618	0.855617	3.674399	3.650277
23.	0.112816	0.957732	427.2246	44.46549	0.087975	0.776689	0.695554	0.959712	0.95848	0.880538	3.057198	3.042049
24.	0.249684	0.935662	140.385	21.62356	0.178402	0.459296	1.446783	0.921228	0.917849	0.662777	5.522337	5.456883
25.	0.157911	0.907568	15.8095	2.900831	0.126187	0.384364	1.29652	0.94164	0.939969	0.542919	3.919333	3.895332

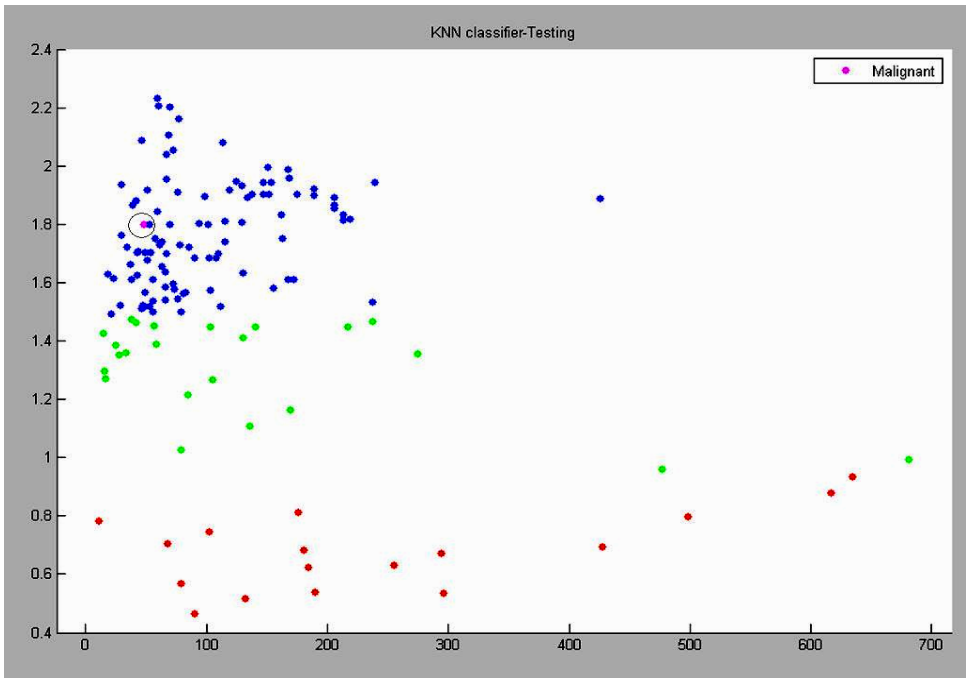

FIGURE 2.3 KNN CLASSIFICATION FOR THE GIVEN QUERY IMAGE

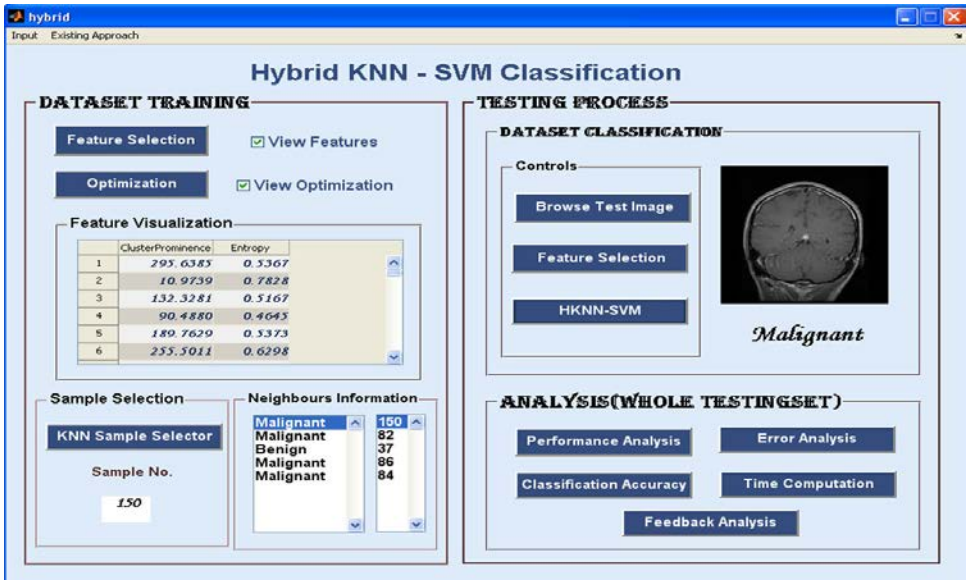

FIGURE 2.4 RESULT OF KNN CLASSIFICATION FOR THE GIVEN QUERY IMAGE

Here, the given query image is classified by the K-NN itself. The result is concluded as "Malignant" category and which is shown in the Figure 2.3. The result is indicated by a circle. Figure 2.4 is indicating the result of the classification.

In SVM, the optimized hyperplane is to be formed in order to split one sample features class from another. The optimized hyperplane will cover the maximum distance margin which will be more accurate classifying feature data tuples than the other hyperplanes. The optimized hyperplane is received using the following expression. The training samples which fall on hyperplane margine lines are called as Support Vectors.

Expression for hyper plane.

$w.t+b = 0$

t – Set of training vectors

w – Vectors perpendicular to the separating hyper plane

b – Offset parameter which allows the increase of the margin

In order to identify the correct maximum margin hyper plane, the SVM tries to maximize the following function with respect to $\vec{w}$ and b:

$$Lp = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t} \propto_i y_i(\vec{w}.\vec{x}_i + b) + \sum_{i=1}^{t} \propto_i \quad (2.2)$$

Where t is the number of training examples, and $\propto_i$, $i = 1, \ldots, t$, are non-negative numbers such that the derivatives of L_P with respect to $\propto_i$ are zero. αi are the Lagrange multipliers and L_P is called the Lagrangian. In this equation, the vectors $\vec{w}$ and constant b define the hyper plane.(Knowl Inf Syst (2008), a survey paper)

Problem Justification

Furthermore, the conventional sensitivity and specificity examination are to be done to justify the result. The two statistical measurements, Sensitivity and specificity are to be used for identifying the performance of a binary classification test. Sensitivity measures the proportion of actual positives which are correctly identified and Specificity measures the proportion of negatives which are correctly identified. Hence, the conventional Partest Graph is plotted based on Cardillo G. (2006), "Clinical test performance", the performance of a clinical test based on the Bayes theorem. The Partest graph helps to know the performance of the execution results. The following constraints are defined for the positive and negative test analysis. The input for the Partest graph function is a 2x2 matrix true Table according to the Table 2.3.

True Positive (TP) - Sick people correctly diagnosed as sick (TP)

False Positive (FP) - Healthy people incorrectly identified as sick (FP)

True Negative (TN)- Healthy people correctly identified as healthy (TN)

False Negative (FN) - Sick people incorrectly identified as healthy (FN)

TABLE 2.3 TRUE TABLE FOR PARTEST GRAPH

	Unhealthy(D+)	**Healthy(D-)**
Positive Test(T+)	True Positives(TP)	False Positives(FP)
Negative Test(T-)	False Negatives(FN)	True Negatives(TN)

Here, the values received are:

True Positive (TP) = 123

True Negative (TN) = 23

False Positive (FP) = 0

False Negative (FN) = 4

Hence, the true Table becomes as follows:

TABLE 2.4 RESULTS ANALYSIS OF TRUE TABLE FOR PARTEST GRAPH

	Unhealthy(D+)	**Healthy(D-)**
Positive Test(T+)	123	0
Negative Test(T-)	4	23

1) Sensitivity

Probability that tests positive on unhealthy subject.

Sensitivity = TP/ (TP+FN)*100 %. (2.3)

2) Specificity

Probability that tests negative on healthy subject

Specificity = TN/ (TN+FP)*100 %. (2.4)

According to the principles of Sensitivity and Specificity, the values of the TP and TN should always be high. The following results have been received by applying Partest test analysis. The execution result of Partest test analysis is shown in the Figure 2.5.

Sensitivity (TP) : 96.9%

False Negative (FN) : 3.1%

Specificity (TN) : 100.0%

False Positive (FP) : 0.0%

Accuracy or Potency : 97.3%

Misclassification Rate : 2.7%

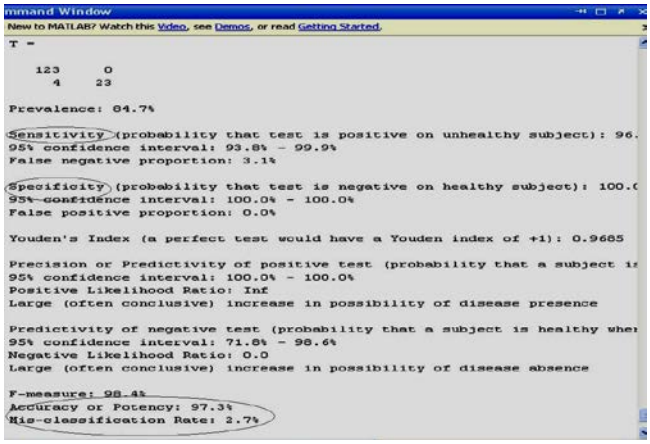

FIGURE 2.5 SENSITIVITY, SPECIFICITY AND ACCURACY JUSTIFICATIONS

The Figure 2.6 shows that the Partest-graph satisfies the result of this classification wherein the rate of FP and FN are highly reduced and the rate of TP and TN are highly increased.

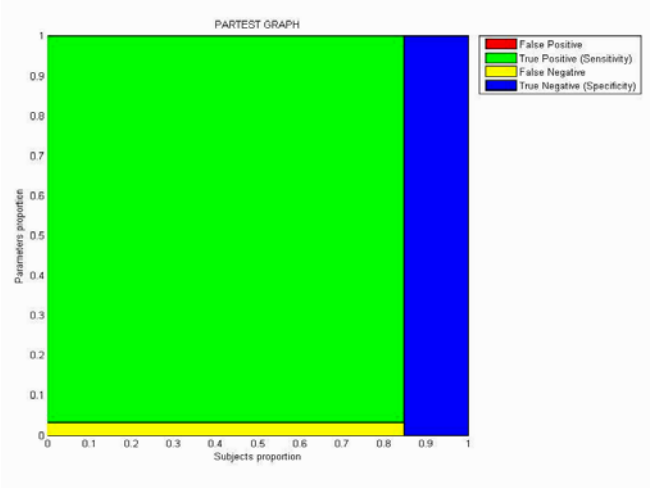

FIGURE 2.6 PARTEST GRAPH FOR SENSITIVITY AND SPECIFICITY ANALYSIS

The Figure 2.7 shows the overall classification accuracy rate of testing set compared to training set. Here, the range of test samples is similar to training sample as mentioned in the Table 4.1. Hence, the result of testing phase deviates slightly compared to training phase. Twenty three Normal MR Images, twenty five Benign MR Images and 102 Malignant MR Images were there during training phase. During testing phase under classification, the 'Normal' is classified as 27, the 'Benign' as 19 and the 'Malignant' as 104 are received. Though the algorithm improves the accuracy of the classification result, it also produces certain misclassification results during its execution. The error rate is calculated at every range of 25 MR images. However, the error rate for every 25 images is very low. The error rate received at each and every twenty images classification processes shown in the Table 2.5 and the Figure 2.8 respectively. The percentage of error rate shows that the misclassification of proposed hybrid algorithm at regular interval. Hence, the accuracy rate is received as 97.3% and misclassification rate as 2.7% as shown in the Figure 2.5.

TABLE 2.5 AVERAGE ERROR RATES AT THE INTERVAL OF 25 IMAGES

No. of Images						
Method	**25**	**50**	**75**	**100**	**125**	**150**
Hybrid	0	0.2400	0	0	0	0

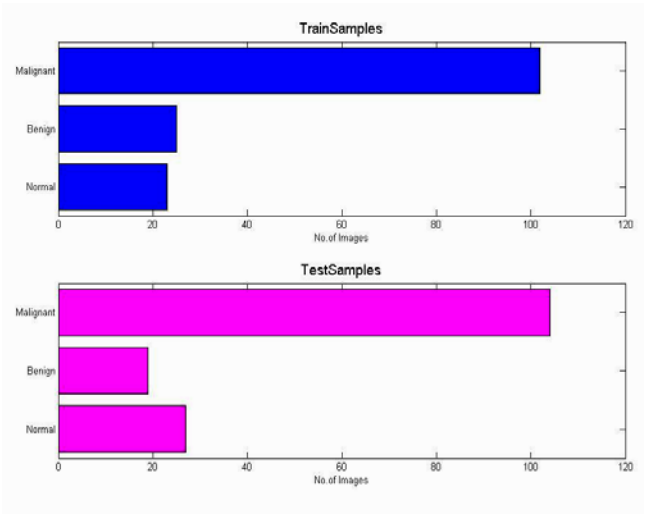

FIGURE 2.7 CLASSIFICATION ACCURACY OF HKNNSVM

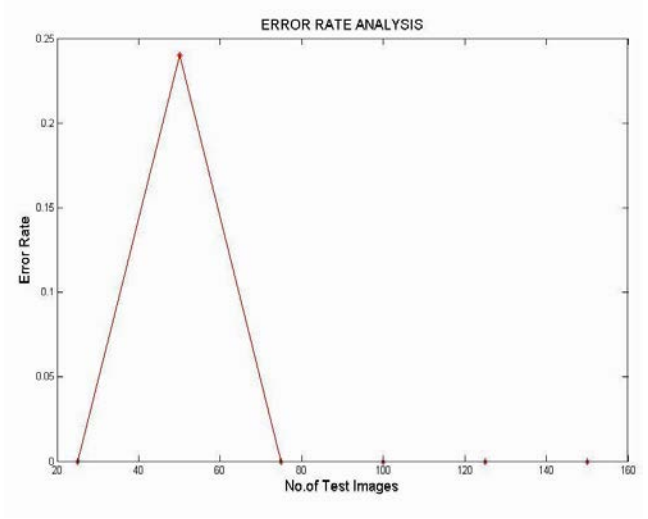

FIGURE 2.8 ERROR RATE ANALYSES

Conclusion

The previous Hybridized KNN and SVM algorithm proposed by Mei et al (2009) aimed to analyze gene expression for cancer classification in binary and multi-class categorization. Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik developed a hybridized SVM and NN algorithm in order to classify general category images including color images. In their algorithm, they were used 'crude' distance calculation to prune the list of neighbors before the most costly accurate distance computation **(**Hao Zhang et al). In this research work, supervised classifications MR Images have been taken into account. About 150 patients' MR images have been used to diagnose the tumour stages such as 'Normal', 'Benign' and 'Malignant' in human brain based on the hybridized KNNSVM classification algorithm. In general, either KNN or SVM will be used to classify the images. But, here, considering their major

drawbacks in certain situations, these two have jointly been applied in this problem. Here, the SVM has also been designed as binary classification processes such as SVM1 and SVM2 in order to classify 'Normal vs Abnormal' and 'Benign vs Malignant'. Hence, the accuracy of the result has been consistently increased and the error rate has been consistently reduced. We have achived 97.3% accuracy and 2.7% misclassification error rate in this paper. In future, a CBIR application is to be proposed by combining the concepts of CBIR and Hybrid KNNSVM algorithm to retrieve similar conditioned MR images based on the given query MR images.

REFERENCES

Cardillo, G. "Clinical test performance", MatLab Central, pp 1-2, 2010.

Durgesh Srivastava, K. and Bhambhu, Lekha "Data Classification Using Support Vector Machine", Journal of Theoretical and Applied Information Technology, pp 1-7, 2009.

Emre Celebi, M., Hitoshi Iyatomi, Joseph Malters, M. and James Grichnik, M. "An Improved Objective Evaluation Measure for Border Detection in Dermoscopy Images", 2009.

John Goodall, R., Conti, Greg and Ma Kwan-liu "Proceedings of the Workshop on Visualization for Computer Security", 2010.

Lam Hong, Lee, Chin Heng, Wan, Tien Fui, Yong and Hui Meian Kok "A Review of Nearest Neighbor-Support Vector Machines Hybrid Classification Models", Universiti Tunku, 2010.

Pettersson, Olle "Implementing LVQ for Age Classification", Master of Science Thesis, Stockholm, Sweden 2007.

Walter, P., Sweeney, J. R., Musavi, Mohamad T. and Guidi, John N. "Classification of Chromosomes Using a Probabilistic Neural Network", University of Maine, Department of Electrical and Computer Engineering, Orono, Maine 04469-5708 (W.P.S., M.T.M.); Received for publication April 22, 1993.

Wu Xindong, Kumar, Vipin, Quinlan, J. Ross, Ghosh, Joydeep, Yang Qiang, Motoda, Hiroshi, McLachlan, Geoffrey J., Ng, Angus, Liu Bing, Yu Philip S., Zhou Zhi-Hua, Steinbach, Michael, Hand, David J., Steinberg, Dan, "Top 10 algorithms in data mining", Knowl Inf Syst DOI 10.1007/s10115-007-0114-2,a survey paper], 2008.